Mathematical Methods in Operations Research 2005
Special issue in honor of *Arie Hordijk*

The $G/M/1$ Queue revisited

Ivo Adan[1], Onno Boxma[1], David Perry[2]

[1] EURANDOM and Department of Mathematics and Computer Science, Eindhoven University of Technology
[2] Department of Statistics, University of Haifa

Received: December 2004

Abstract The $G/M/1$ queue is one of the classical models of queueing theory. The goal of this paper is two-fold: (i) To introduce new derivations of some well-known results, and (ii) to present some new results for the $G/M/1$ queue and its variants. In particular, we pay attention to the $G/M/1$ queue with a set-up time at the start of each busy period, and the $G/M/1$ queue with exceptional first service.

For Arie Hordijk on his 65-th birthday, in friendship and admiration

1 Introduction

The $G/M/1$ queue is one of the classical models of queueing theory. The goal of this paper is two-fold: (i) To introduce new derivations of some well-known results, and (ii) to present some new results for the $G/M/1$ queue and its variants. In particular, we pay attention to the $G/M/1$ queue with a set-up time at the start of each busy period, to the $G/M/1$ queue with exceptional first service time, and to the cycle maximum of the $G/M/1$ queue. The main methods in the paper are (i) martingale techniques, (ii) transform techniques, and (iii) sample-path arguments, exploiting duality between the attained and virtual waiting time processes.

Treatments of the $G/M/1$ queue may be found in several books on queueing theory; see, e.g., Asmussen [1], Cohen [2], Prabhu [9] and Takács [11]. Doshi [3] has studied the $GI/G/1$ queue with vacations or set-up times. The decomposition result that he obtains for the waiting time distribution is quite involved in the case of set-up times; in the case of exponential service times and phase-type set-up times, we obtain more explicit decomposition results.

The paper is organized as follows. Below we describe the model and introduce some notation. Section 2 introduces the *attained waiting time* process of the $G/M/1$ queue and relates it to the *virtual waiting time* process (or work process) of that same queue. In Section 3 the *attained waiting time* is shown to be exponentially distributed. A brief derivation of the idle period distribution is presented in Section 4, using a martingale approach. Sections 5 and 6 are devoted to the $G/M/1$ queue with set-up times. We derive a decomposition result for the attained waiting time process, thus also retrieving a sojourn time decomposition result of Doshi [3]. Like in the case without set-up times, we use a martingale to derive an expression for the Laplace-Stieltjes transform of the idle period distribution. Section 6 considers the case of Erlang set-up times. In Section 7 we study the $G/M/1$ queue with exceptional first service time in a busy period. We obtain the joint distribution of the busy and idle period. For the case of the ordinary $G/M/1$ queue, a known result (cf. [9]) is re-derived. The last section of the paper is devoted to a study of the cycle maximum in a busy period of the $G/M/1$ queue. An approach based on the *attained waiting time process* is chosen for the steady-state case. In the case of overload, an approach based on the *virtual waiting time process* is employed to analyze the cycle maximum, given that the busy period is finite.

2 The $G/M/1$ Queue

We consider the classical $G/M/1$ queue. The times between successive arrivals are i.i.d. random variables $S_1, S_2, \ldots$, with distribution $G(\cdot)$, Laplace-Stieltjes transform (LST) $G^*(\alpha)$ and mean $1/\lambda$. The service requirements of the arriving customers are i.i.d. random variables $Z_1, Z_2, \ldots$, which are exponentially distributed with mean $1/\mu$. All interarrival and service times are assumed to be independent. Service is in order of arrival. The traffic load is denoted by $\rho := \lambda/\mu$. It is assumed that $\rho < 1$ (unless stated otherwise).

Several derivations in this study are based on the sample path analysis of two *dual* compound processes; the so-called *virtual waiting time* (VWT) and the *attained waiting time* (AWT) processes. Formally, let $\mathbf{N} = \{N(t) : t \geq 0\}$ and $\mathbf{\Lambda} = \{\Lambda(t) : t \geq 0\}$ be counting processes such that for all $t \geq 0, n = 0, 1, ...$ and $m = 0, 1, ...$: $\{N(t) \geq n\} = \{Z_1 + ... + Z_n \leq t\}$ and $\{\Lambda(t) \geq m\} = \{S_1 + ... + S_m \leq t\}$. Obviously, $\mathbf{N}$ is a Poisson process with rate μ and $\mathbf{\Lambda}$ is a renewal process whose inter-renewal distribution is $G(\cdot)$ with mean $1/\lambda$. Now define the continuous time random walk $\mathbf{X} = \{X(t) : t \geq 0\}$ such that $X(t) = t - (S_1 + ... + S_{N(t)})$ and $\mathbf{Y} = \{Y(t) : t \geq 0\}$ such that $Y(t) = (Z_1 + ... + Z_{\Lambda(t)}) - t$. Then construct the reflected processes $\mathbf{A} = \{A(t) : t \geq 0\}$ and $\mathbf{V} = \{V(t) : t \geq 0\}$, respectively, by

$$A(t) = X(t) - \min_{0 \leq s < t} X(s) \quad \text{and} \quad V(t) = Y(t) - \min_{0 \leq s < t} Y(s).$$

Here $\mathbf{A}$ is interpreted as the conditional AWT process of the $G/M/1$ queue in which the idle periods are deleted and the busy periods are glued together.

The process $\mathbf{V}$ is interpreted as the VWT process (or the *work* process) of the same $G/M/1$ queue. The processes $\mathbf{V}$ and $\mathbf{A}$ are dual processes with respect to waiting times. While $V(t)$ is interpreted as the time a customer would have to wait in line if he arrived at t, $A(t)$ is interpreted as the time already attained (or elapsed) since the arrival of the customer being served at t. In other words, while $\mathbf{V}$ designates the waiting time of a virtual customer by looking forward in time (the customer is virtual in the sense that he did not arrive at t and thus, in practice, he contributed nothing to the work), $\mathbf{A}$ designates the waiting time of a real customer by looking backward in time. As a result, while the VWT might be sometimes equal to 0, the AWT cannot be 0 because the served customer "sees" at least himself in the system. By that interpretation, the steady state law of $\mathbf{A}$ and that of $\mathbf{V}$ must be closely related to each other. In fact, it can be shown by construction (see, e.g., Perry et al. [6]) that the steady state law of $\mathbf{A}$ is equal to that of the conditional steady state law of $\mathbf{V}$ given that the idle periods (the time periods in which $\mathbf{V} = 0$) are deleted and the busy periods are glued together. Note that $\rho < 1$ implies that both $X(t)$ and $Y(t)$ tend to $-\infty$ *a.s.*, so that $\mathbf{A}$ and $\mathbf{V}$ are regenerative processes. Furthermore, the cycles associated with $\mathbf{A}$ are the busy periods and those associated with $\mathbf{V}$ are the busy cycles (the busy cycle is composed of busy period plus idle period). Also, it can be shown (see, e.g., Perry et al. [6]) that the stopping times $T = \inf\{t \geq 0 : X(t) \leq 0\}$ and $\tau = \inf\{t \geq 0 : Y(t) = 0\}$ are the same random variables that represent the busy period of the same $G/M/1$ queue.

Remark 1 A busy cycle generated by $\mathbf{V}$ is $C = \inf\{t \geq \tau : V(t) > 0\}$. Then, $C - T$ and $-A(T)$ are also the same random variables that represent the idle period. Also, while the sample path of $\mathbf{V}$ is continuous at τ and $V(\tau-) = V(\tau+) = 0$, $A(T)$ is a point of discontinuity since by definition $A(T-) > 0 > A(T) < A(T+) = 0$.

3 Density of the Attained Waiting Time

We first study the AWT process $\mathbf{A}$, showing that its steady state distribution is exponential. We define the steady state random variable $A = \lim_{t\to\infty} A(t)$, where the latter limit is defined in terms of weak convergence. Let $f_A(\cdot)$ be the equilibrium density of $\mathbf{A}$. A *level-crossings* argument shows that it satisfies the following steady state equation:

$$f_A(x) = \mu \int_x^\infty [1 - G(w - x)] f_A(w) \mathrm{d}w. \tag{1}$$

Rewrite this equation into:

$$f_A(x) = \mu \int_0^\infty [1 - G(y)] f_A(y + x) \mathrm{d}y. \tag{2}$$

Differentiate to get

$$f'_A(x) = \mu \int_0^\infty [1 - G(y)] f'_A(y + x)\mathrm{d}y,$$

where $f'_A(\cdot)$ is the derivative with respect to $f_A(\cdot)$. We know that, for $\rho < 1$, A has a unique density. Noticing that $f_A(x)$ and $f'_A(x)$ satisfy the same equation, it follows that $f'_A(x)$ equals $f_A(x)$, up to a multiplicative constant. Solving $f'_A(x) = \eta f_A(x)$ with $\int_0^\infty f_A(x)\mathrm{d}x = 1$ yields

$$f_A(x) = \eta e^{-\eta x}, \qquad x > 0. \tag{3}$$

Here η is implicitly defined as the unique solution, in $(0, \mu)$, of

$$\eta = \mu[1 - G^*(\eta)]. \tag{4}$$

The fact that η satisfies (4) follows by substitution of (3) in (2). The uniqueness statement follows since $G^*(0) = 1$, $G^*(\infty) = 0$ and $G^*(\alpha)$ is a monotone decreasing convex function, combined with $\rho < 1$ (which implies that the derivative of the right-hand side of (4) is $1/\rho > 1$). We conclude that the steady state law of the process $\mathbf{A}$, i.e., the AWT process of the $G/M/1$ queue in which the idle periods are deleted, is $\exp(\eta)$.

Remark 2 The last result implies that the sojourn times of the $G/M/1$ queue are also *exp*(η) distributed (the latter statement is a well-known result, see [2] or [5]). To see this, note that the sojourn times are the peak values of the AWT process. But these peak values occur at the arrival instants of the Poisson process $\mathbf{N}$. Hence, by PASTA, the limiting distribution of the peak values of the AWT process equals the stationary distribution.

4 Martingale Approach for the Idle Period

We now turn to the idle period. In the $G/M/1$ queue, the busy period and the idle period are not necessarily independent. Just for the sake of convenience, the analysis is based on the *random walk* $\hat{\mathbf{X}} := -\mathbf{X}$. Consider the process $\mathbf{M} = \{M(t) : t \geq 0\}$, where

$$M(t) = \varphi(\alpha) \int_0^t \mathrm{e}^{-\alpha \hat{X}(s)}\mathrm{d}s + \mathrm{e}^{-\alpha \hat{X}(0)} - e^{-\alpha \hat{X}(t)}, \tag{5}$$

and $\varphi(\alpha) := \alpha - \mu[1 - G^*(\alpha)]$ is the *exponent* of $\hat{\mathbf{X}}$. It is well-known that $\mathbf{M}$ is a martingale (see Kella and Whitt [4]), and by applying the optional sampling theorem for T (clearly, T is the same for both $\hat{\mathbf{X}}$ and $\mathbf{X}$) to the martingale $\mathbf{M}$ we see that $EM(T) = 0$, thus obtaining the *fundamental identity* (with the substitution $\hat{X}(0) = 0$)

$$\varphi(\alpha) E\left(\int_0^T \mathrm{e}^{-\alpha \hat{X}(s)}\mathrm{d}s\right) = -1 + E(\mathrm{e}^{-\alpha \hat{X}(T)}). \tag{6}$$

We now can prove the following result (see also, e.g., p. 36 of Prabhu [9]).

Theorem 1 *The LST of the idle period of the $G/M/1$ queue is given by*

$$Ee^{-\alpha I} = \frac{\eta - \mu[1 - G^*(\alpha)]}{\eta - \alpha}.$$

Proof The fact that $\int_0^T \mathrm{e}^{-\alpha \hat{X}(s)} \mathrm{d}s = \int_0^T \mathrm{e}^{-\alpha \hat{A}(s)} \mathrm{d}s$ (where $\hat{\mathbf{A}} := -\mathbf{A}$), follows immediately by the definition of T. We thus express (6) in terms of the process $\mathbf{A}$:

$$\varphi(\alpha) E\left(\int_0^T \mathrm{e}^{\alpha A(s)} \mathrm{d}s\right) = -1 + E(\mathrm{e}^{-\alpha \hat{A}(T)}). \tag{7}$$

Using the theory of regenerative processes and the fact that $-A(T) = \hat{A}(T) = I$, the idle period, we obtain:

$$E(\mathrm{e}^{-\alpha I}) = 1 + \varphi(\alpha) ETE(\mathrm{e}^{\alpha A}). \tag{8}$$

Now use the fact that A is $\exp(\eta)$ distributed (cf. (3)), and that $ET = 1/\eta$. To see the latter result, note that level 0 is down-crossed by $\mathbf{A}$ (alternatively, up-crossed by $\hat{\mathbf{A}}$) exactly once during the cycle T (the down-crossing occurs at T since $A(T-) > 0 > A(T)$). By *level crossing theory*, $f(0)$ is the rate of the long-run average number of down-crossings of level 0. Thus, $ET = 1/f(0) = 1/\eta$.

Remark 3 Of course, the LST of the idle period may also be obtained directly from Lindley's equation (see, e.g., Asmussen [1]),

$$I = (S - A | S - A > 0),$$

where the generic random variable S denotes the inter-arrival time and A the sojourn time (which is $\exp(\eta)$).

5 Constant Set-Up Times

We now turn our attention to a $G/M/1$ queue with a set-up time, R, at the beginning of each busy period. It appears to be convenient to start with the case of a deterministic set-up time $R = x$. Subsequently, for general set-up times, the results for the constant case may be integrated w.r.t. the distribution of the set-up time. The expressions obtained for general set-up times, however, turn out to be not very explicit. Consequently, the special case of an Erlang-distributed set-up time will be discussed separately in Section 6 (and yields more explicit results).

Consider the process $\mathbf{A_x}$=$\{A_x(t) : t \geq 0\}$, where

$$A_x(t) = x + X(t) - \min_{0 \leq s < t} (x + X(s)),$$

with x some nonnegative constant. The process $\mathbf{A_x}$ can be visualized as process $\mathbf{A}$ that is lifted during each busy period of the original process. During the first busy period of the corresponding $G/M/1$ queue *without* set-up time, it is lifted to level x, during the second one to level $x - I_1$, where I_1 is the first idle period, during the third one to level $x - I_1 - I_2$ and so on. Define $K = \inf\{k : x - I_1 - I_2 - \cdots - I_k < 0\}$. That is, K is the number of negative ladder heights during the busy period $T_x = \inf\{t : A_x(t) \leq 0\}$. Let $\mathbf{L} = \{L(t) : t \geq 0\}$ denote the level process at which $\mathbf{A}$ is lifted at time t, i.e., $L(t)$ is the value with which $A(t)$ is lifted, and let A_x, L and A denote the steady-state random variables associated with the processes $\mathbf{A_x}$, $\mathbf{L}$ and $\mathbf{A}$, respectively (recall from (3) that A is $\exp(\eta)$). At time $t \geq 0$, $L(t)$ does not depend on the current busy period of $A(t)$, but only on the ones prior to the current one. Hence, $L(t)$ and $A(t)$ are independent, and thus (by letting t tend to infinity) we can conclude that:

Theorem 2 *$A_x \stackrel{\mathcal{D}}{=} A + L$, where A and L are independent.*

Remark 4 Theorem 2 is also valid for general set-up times. In fact, it implies that the steady-state sojourn time in the $G/M/1$ with set-up times can be decomposed as the sum of two independent random variables: the steady-state sojourn time in the $G/M/1$ *without* set-up times and the steady-state level of lifting (cf. Remark 2).

To determine the steady-state distribution of $\mathbf{A_x}$, we need to determine the distribution of $\mathbf{L}$. The expected number of busy periods in a cycle of $\mathbf{L}$ that are lifted higher than y is equal to $1 + m(x - y)$, where $m(t)$ denotes the renewal function of the process of idle periods $\{I_n\}$. Hence, we have

$$\Pr(L > y) = \frac{1 + m(x - y)}{1 + m(x)}, \qquad 0 \leq y < x, \qquad \Pr(L = x) = \frac{1}{1 + m(x)}. \tag{9}$$

Note that L has probability mass at x. For example, in case that the inter-arrival times S_i are also exponential with parameter λ, we have $m(x) = \lambda x$, and then

$$\Pr(L > y) = 1 - \frac{\lambda y}{1 + \lambda x}, \qquad 0 \leq y < x, \qquad \Pr(L = x) = \frac{1}{1 + \lambda x}.$$

The steady state law of $\mathbf{A}_x$ is introduced in the following lemma.

Lemma 1

$$Ee^{-\alpha A_x} = \frac{\eta}{\eta + \alpha} \left[\frac{e^{-\alpha x} * m(x) + e^{-\alpha x}}{1 + m(x)} \right],$$

where "$*$" is the convolution sign.

Proof By Theorem 2,

$$Ee^{-\alpha A_x} = Ee^{-\alpha A} Ee^{-\alpha L},$$

where from (3):

$$Ee^{-\alpha A} = \frac{\eta}{\eta + \alpha}.$$

Also, by (9),

$$Ee^{-\alpha L} = \int_{y=0}^{x} e^{-\alpha y} \mathrm{d}\Pr(L < y) + e^{-\alpha x} \Pr(L = x) = \frac{e^{-\alpha x} * m(x) + e^{-\alpha x}}{1 + m(x)},$$

which completes the proof.

Remark 5 If the set-up time R has a general distribution, then one can easily get the LST of the steady-state distribution of the AWT process by integrating the LST of A_x w.r.t. the set-up time distribution. Let the random variable A_R denote the AWT in steady-state. Then

$$Ee^{-\alpha A_R} = \frac{\eta}{\eta + \alpha} \int_0^\infty \frac{e^{-\alpha x} + \mathrm{e}^{-\alpha x} * m(x)}{1 + m(x)} \mathrm{d}\Pr(R \le x). \tag{10}$$

It should be observed, though, that the expression involves the renewal function $m(\cdot)$ of the idle periods of the $G/M/1$ queue.

We now proceed to study the idle and busy period. Let I_x and T_x be the idle period and busy period, respectively, associated with $\mathbf{A}_x$.

Lemma 2

$$Ee^{-\alpha I_x} = \frac{\alpha - \mu(1 - G^*(\alpha))}{ET_x} \frac{\eta}{\eta - \alpha} \left[\frac{e^{\alpha x} * m(x) + e^{\alpha x}}{1 + m(x)} \right] + \mathrm{e}^{\alpha x},$$

with

$$ET_x = \mu[\int_0^\infty (1 - G(u)) dF_{A_x}(u)]^{-1}, \tag{11}$$

and where $F_{A_x}(\cdot)$ is the distribution whose LST is given in Lemma 1.

Proof Consider the process $\tilde{\mathbf{M}} = \{\tilde{M}(t) : t \ge 0\}$ where

$$\tilde{M}(t) = \tilde{\varphi}(\alpha) \int_0^t \mathrm{e}^{-\alpha(x + X(s))} \mathrm{d}s + \mathrm{e}^{-\alpha x} - e^{-\alpha(x + X(t))}$$

and

$$\tilde{\varphi}(\alpha) = -[\alpha + \mu(1 - G^*(-\alpha))].$$

It is readily seen that $\tilde{\mathbf{M}}$ is a martingale and by applying the optional sampling theorem for $T_x = \inf\{t : A_x(t) \le 0\}$ to the martingale $\tilde{\mathbf{M}}$ we see that

$$\tilde{\varphi}(\alpha) E \left(\int_0^{T_x} \mathrm{e}^{-\alpha A_x(s)} \mathrm{d}s \right) = -\mathrm{e}^{-\alpha x} + E(\mathrm{e}^{-\alpha A_x(T_x)}). \tag{12}$$

By the theory of regenerative processes, the left hand side of (12) is

$$\frac{\tilde{\varphi}(\alpha) E \mathrm{e}^{-\alpha A_x}}{ET_x}.$$

Also, $-A_x(T_x)$ can be interpreted as the idle period I_x. Thus from (12),

$$\begin{aligned} E(\mathrm{e}^{\alpha I_x}) &= \tfrac{\tilde{\varphi}(\alpha) E \mathrm{e}^{-\alpha A_x}}{ET_x} + \mathrm{e}^{-\alpha x} \\ &= \tfrac{\tilde{\varphi}(\alpha)}{ET_x} \tfrac{\eta}{\eta+\alpha} \left[\tfrac{e^{-\alpha x} * m(x) + e^{-\alpha x}}{1+m(x)} \right] + \mathrm{e}^{-\alpha x}, \end{aligned} \tag{13}$$

where the second step follows by Lemma 1. Finally, ET_x is the reciprocal of the rate of down-crossings of level 0 by $\mathbf{A}_x$. Thus, by *level crossing theory*, (11) follows. Now replace α by $-\alpha$ in (13) and the result follows.

Remark 6 In order to obtain the LST of the idle period associated with a generally distributed set-up time R, we apply the law of total probability in (13) to get

$$E(\mathrm{e}^{-\alpha I}) = \int_0^\infty E(\mathrm{e}^{-\alpha I_x}) \mathrm{d} \Pr(R \leq x).$$

6 Erlang Set-up Times

In this section we consider the special case of Erlang distributed set-up times, i.e., the set-up time R is the sum of n exponentials with parameter ν. Then the AWT process $\mathbf{A_R}$ can be visualized as process $\mathbf{A}$ that is lifted during each busy period by at least one and at most n exponentials. Let L_m denote the number of exponentials lifting $\mathbf{A}$ during the mth busy period. Clearly, $\mathbf{L} = \{L_m, m = 0, 1, 2, \ldots\}$ is a Markov chain with states $\{1, \ldots, n\}$ and the one-step transition probabilities $p_{i,j}$ are given by

$$\begin{aligned} p_{i,i-k} &= \Pr(X_1 + \cdots + X_k < I < X_1 + \cdots + X_{k+1}) \\ &= \frac{(-\nu)^k}{k!} I^{*(n)}(\nu), \qquad k = 0, 1, 2, \ldots, i-1; \\ p_{i,n} &= 1 - p_{i,i} - p_{i,i-1} - \cdots - p_{i,1}, \end{aligned}$$

where $X_1, X_2, \ldots$ are independent exponentials, each with parameter ν, and $I^{*(k)}(\cdot)$ is the kth derivative of the LST of the idle period I associated with the $G/M/1$ without set-up times (see Section 4). Here we used that

$$\begin{aligned} \Pr(I < X_1 + \cdots + X_{k+1}) &= \int_0^\infty e^{-\nu x} \sum_{i=0}^k \frac{(\nu x)^i}{i!} \mathrm{d} \Pr(I \leq x) \\ &= \sum_{i=0}^k \frac{\nu^i}{i!} \int_0^\infty e^{-\nu x} x^i \mathrm{d} \Pr(I \leq x) \\ &= \sum_{i=0}^k \frac{(-\nu)^i}{i!} I^{*(i)}(\nu). \end{aligned}$$

Let $\pi_1, \ldots, \pi_n$ denote the steady-state probabilities of $\mathbf{L}$. These probabilities can be easily calculated recursively: Let v_k be the expected number of visits to state k till the first return to state n, when starting in state n, so $v_n = 1$ and

$$v_k = \sum_{l=k+1}^{n} v_l p_{l,k}, \qquad k = n-1, n-2, \ldots, 1.$$

Then the steady-state probabilities follow from normalization, i.e.,

$$\pi_k = \frac{v_k}{v_1 + \cdots + v_n}, \qquad k = 1, \ldots, n.$$

Hence, we have (see Theorem 2 and Remark 4),

$$A_R \stackrel{\mathcal{D}}{=} A + L,$$

where

$$L \stackrel{\mathcal{D}}{=} \begin{cases} X_1, & \text{w.p. } \pi_1, \\ X_1 + X_2, & \text{w.p. } \pi_2, \\ \vdots & \\ X_1 + \cdots + X_n, & \text{w.p. } \pi_n. \end{cases}$$

Remark 7 The above result can be easily extended to mixed Erlang set-up times. Suppose that, with probability p_i, $i = 1, \ldots, n$, the set-up time R is the sum of i independent exponentials, each with parameter ν. Then the steady-state distribution of $\mathbf{L}$ is given by

$$\pi_i = \frac{\sum_{k=1}^{n} p_k \pi_i^k / \pi_k^k}{\sum_{k=1}^{n} p_k / \pi_k^k}, \qquad i = 1, \ldots, n,$$

where $\pi_1^k, \ldots, \pi_k^k$ denote the steady-state probabilities for Erlang-k distributed set-up times, with parameter ν.

7 Joint Distribution of Busy and Idle Period

In this section we determine the LST of the joint distribution of the busy period T and idle period I in the $G/M/1$ queue, for the case that the *first* service time Z_1 of the busy period is x. By integrating the result w.r.t. the probability distribution of Z_1, we subsequently also determine the LST of the joint distribution of the busy period and idle period in the $G/M/1$ queue with either set-up time or an exceptional first service time. Introduce, for $\mathrm{Re}\ \alpha_1, \alpha_2 \geq 0$, $x \geq 0$:

$$k(x, \alpha_1, \alpha_2) := E(\mathrm{e}^{-\alpha_1 T - \alpha_2 I} | Z_1 = x), \tag{14}$$

$$K(s, \alpha_1, \alpha_2) := \int_0^\infty \mathrm{e}^{-sx} k(x, \alpha_1, \alpha_2) \mathrm{d}x. \tag{15}$$

Also introduce $\hat{s} = \hat{s}(\alpha_1)$, the unique zero of $1 - \frac{\mu}{\mu - s} G^*(\alpha_1 + s)$ in the righthalf α_1-plane (see, e.g., Cohen [2], p. 226).

Theorem 3 *For* Re $s, \alpha_1, \alpha_2 \geq 0$,

$$K(s,\alpha_1,\alpha_2) = \frac{\mu - s}{\mu - s - \mu G^*(\alpha_1+s)}[\frac{G^*(\alpha_1+s)-G^*(\alpha_2)}{\alpha_2-\alpha_1-s} - \frac{\mu}{\mu-s}G^*(\alpha_1+s)\frac{G^*(\alpha_1+\hat{s})-G^*(\alpha_2)}{\alpha_2-\alpha_1-\hat{s}}]. \quad (16)$$

Proof Conditioning on the two possibilities that the first interarrival time $S_1 \geq x$ and $S_1 < x$, we can write:

$$k(x,\alpha_1,\alpha_2) = \int_{t=x}^{\infty} \mathrm{e}^{-\alpha_1 x}\mathrm{e}^{-\alpha_2(t-x)}\mathrm{d}G(t) + \int_{z=0}^{\infty}\mu\mathrm{e}^{-\mu z}\int_{t=0}^{x}\mathrm{e}^{-\alpha_1 t}k(x-t+z,\alpha_1,\alpha_2)\mathrm{d}G(t)\mathrm{d}z. \quad (17)$$

Taking the LT (Laplace Transform) w.r.t. x and changing integration orders yields:

$$K(s,\alpha_1,\alpha_2) = \int_{x=0}^{\infty}\mathrm{e}^{-(\alpha_1+s)x}\int_{u=0}^{\infty}\mathrm{e}^{-\alpha_2 u}\mathrm{d}G(x+u)\mathrm{d}x + \mu\int_{t=0}^{\infty}\mathrm{e}^{-(\alpha_1+s)t}\mathrm{d}G(t)\int_{u=0}^{\infty}\int_{z=0}^{\infty}\mathrm{e}^{-su}\mathrm{e}^{-\mu z}k(u+z,\alpha_1,\alpha_2)\mathrm{d}u\mathrm{d}z$$
$$= \frac{G^*(\alpha_1+s)-G^*(\alpha_2)}{\alpha_2-\alpha_1-s} + \mu G^*(\alpha_1+s)\frac{K(s,\alpha_1,\alpha_2)-K(\mu,\alpha_1,\alpha_2)}{\mu-s}. \quad (18)$$

Hence

$$K(s,\alpha_1,\alpha_2)[1-\frac{\mu}{\mu-s}G^*(\alpha_1+s)] = \frac{G^*(\alpha_1+s)-G^*(\alpha_2)}{\alpha_2-\alpha_1-s} - \frac{\mu}{\mu-s}G^*(\alpha_1+s)K(\mu,\alpha_1,\alpha_2). \quad (19)$$

It remains to determine $K(\mu,\alpha_1,\alpha_2)$. A standard analyticity argument gives (remember the definition of $\hat{s}$ above):

$$K(\mu,\alpha_1,\alpha_2) = \frac{G^*(\alpha_1+\hat{s})-G^*(\alpha_2)}{\alpha_2-\alpha_1-\hat{s}}, \quad \text{Re } \alpha_1,\alpha_2 \geq 0. \quad (20)$$

Substitution in (19) finally gives the statement of the theorem.

Remark 8 Determination of $k(x,\alpha_1,\alpha_2)$.
In principle, one can invert $K(s,\alpha_1,\alpha_2)$ to obtain $k(x,\alpha_1,\alpha_2)$. Rewrite (16) as follows (s should be such that the sum converges):

$$K(s,\alpha_1,\alpha_2) = \sum_{j=0}^{\infty}(\frac{\mu}{\mu-s})^j(G^*(\alpha_1+s))^j[\frac{G^*(\alpha_1+s)-G^*(\alpha_2)}{\alpha_2-\alpha_1-s} - \frac{\mu}{\mu-s}G^*(\alpha_1+s)\frac{G^*(\alpha_1+\hat{s})-G^*(\alpha_2)}{\alpha_2-\alpha_1-\hat{s}}]. \quad (21)$$

Now observe that

$$G^*(\alpha_1 + s) = G^*(\alpha_1) \int_{x=0}^{\infty} \mathrm{e}^{-sx} [\frac{\mathrm{e}^{-\alpha_1 x} \mathrm{d}G(x)}{\int_0^\infty \mathrm{e}^{-\alpha_1 y} \mathrm{d}G(y)}],$$

which equals the product of $G^*(\alpha_1)$ and the LST of $\Pr(S_1 < x | S_1 < E_1)$, with E_1 exponentially distributed with mean $1/\alpha_1$. The first part of the righthand side of (21) can now be inverted term by term (care should be taken of the fact that $K(s, \alpha_1, \alpha_2)$ is an LT, and not an LST, w.r.t. x). The term between square brackets in (21) is easier to invert; notice that the first term in the righthand side of (17) is the inverse of $(G^*(\alpha_1 + s) - G^*(\alpha_2))/(\alpha_2 - \alpha_1 - s)$. Of course, for specific choices of the interarrival time distribution, like the Erlang distribution, it is a rather straightforward task to obtain $k(x, \alpha_1, \alpha_2)$ by inversion of the expression in (16).

Remark 9 The LST of the joint distribution of busy and idle period.
It should be noted that $\mu K(\mu, \alpha_1, \alpha_2)$ is the LST of the joint distribution of the busy period T and idle period I in the ordinary $G/M/1$ queue, in which *also* the first service time Z_1 is $\exp(\mu)$ distributed. The result agrees with Formula (47) on p. 57 of Prabhu [10]. Taking $\alpha_1 = \alpha_2$ yields the LST of the busy cycle length in the $G/M/1$ queue. Next suppose that the first service time Z_1 is hyperexponentially distributed, with density $\sum_{i=1}^k p_i \nu_i \mathrm{e}^{-\nu_i x}$. In that case,

$$E(\mathrm{e}^{-\alpha_1 T - \alpha_2 I}) = \sum_{i=1}^{k} p_i \nu_i K(\nu_i, \alpha_1, \alpha_2). \qquad (22)$$

Finally suppose that the first service time Z_1 is Erlang-k distributed, with parameter ν. Then it is easily verified that

$$E(\mathrm{e}^{-\alpha_1 T - \alpha_2 I}) = \frac{(-1)^{k-1} \nu^k}{(k-1)!} \frac{d^{k-1}}{ds^{k-1}} K(s, \alpha_1, \alpha_2)|_{s=\nu}. \qquad (23)$$

8 Cycle Maximum

In this section we introduce two approaches for analysis of the cycle maximum of the $G/M/1$ queue; the first is based on the AWT, the second on the VWT. We refer to Cohen [2], Section III.7.5, for an expression for this cycle maximum in the form of a contour integral. That is a result for $\rho \leq 1$. In Subsection 8.2 we consider the case $\rho > 1$.

8.1 AWT Approach

Recall that $X(t) = t - (S_1 + S_2 + ... + S_{\Lambda(t)})$ and $T = \inf\{t : X(t) \leq 0\}$. Let $M = \max_{0 \leq t \leq T} X(t)$. In this section we compute the law of M, the *cycle maximum* of the busy cycle.

Theorem 4

$$\Pr(M > x) = \frac{e^{-\eta x}(1 - Ee^{-\eta I})}{1 - e^{-\eta x} Ee^{-\eta I_x}},$$

where $Ee^{-\eta I}$ is given in Theorem 1 and $Ee^{-\eta I_x}$ is given in Lemma 2 above with α replacing η.

Proof

$$\begin{aligned}
\Pr(A > x) &= \Pr(\max_{0 \le t < \infty} X(t) > x) \\
&= \Pr(\{\max_{0 \le t < T} X(t) > x\} \cup \{\max_{T \le t < \infty} X(t) > x\}) \\
&= \Pr(\max_{0 \le t < T} X(t) > x) + \Pr(\max_{T \le t < \infty} X(t) > x) \\
&\quad - \Pr(\{\max_{0 \le t < T} X(t) > x\} \cap \{\max_{T \le t < \infty} X(t) > x\}) \\
&= \Pr(M > x) + \Pr(\max_{0 \le t < \infty} X(t) > x + I) \\
&\quad - \Pr(\max_{T \le t < \infty} X(t) > x \mid \max_{0 \le t < T} X(t) > x) \Pr(\max_{0 \le t < T} X(t) > x) \\
&= \Pr(M > x) + \Pr(A > x + I) \\
&\quad - \Pr(\max_{T \le t < \infty} X(t) > x \mid \max_{0 \le t < T} X(t) > x] \Pr[M > x).
\end{aligned} \tag{24}$$

Define the stopping time $T_x = \inf\{t : X(t) = x\}$. Given the event $\{\max_{0 \le t < T} X(t) > x\}$ occurred, it follows by the strong Markov property at T_x that

$$\begin{aligned}
\Pr(\max_{T \le t < \infty} X(t) > x \mid \max_{0 \le t < T} X(t) > x) &= \Pr(\max_{0 \le t < \infty} X(t) > x + I_x) \\
&= \Pr(A > x + I_x).
\end{aligned}$$

Thus, we obtain in (24):

$$e^{-\eta x} = \Pr(M > x) + e^{-\eta x} Ee^{-\eta I} - \Pr(M > x) e^{-\eta x} Ee^{-\eta I_x},$$

and the theorem follows.

8.2 The Case $\rho > 1$; VWT Approach

Recall that the S_i are i.i.d., random variables with distribution $G(\cdot)$ and mean $1/\lambda$. Similarly, the Z_j are i.i.d., random variables such that $Z_j \backsim \exp(\mu)$. Also, $\mathbf{N}$ is a Poisson process with rate μ and $\mathbf{\Lambda}$ is a renewal process with interrenewal mean $1/\lambda$. In the present subsection we assume that $\rho := (\lambda/\mu) > 1$ and study the cycle maximum in a busy period of the overloaded $G/M/1$ queue. Indeed, the distribution of the maximum is improper. However, the conditional distribution of the cycle maximum given that the busy period is finite is a proper distribution. Formally, let $Z \backsim \exp(\mu)$ be a random variable independent of the process $\mathbf{Y}$ (recall that $Y(t) = (Z_1 + ... + Z_{A(t)}) - t)$ and define the stopping times

$$T_Z^- = \inf\{t : Y(t) = -Z\}$$

and

$$T_Z^+ = \inf\{t : Y(t) \geq 0\}.$$

Note that T_Z^- can be interpreted as the busy period of the $G/M/1$ queue (with inter-arrival distribution $G(\cdot)$ and service rate μ) and the random variable

$$M = \max_{0 \leq t \leq T_Z^-} (Z + Y(t))$$

is the cycle maximum.

To compute the law of M we use the following argument. Let t be a record time for $\{Z+Y(t) : 0 \leq t \leq T_Z^-\}$ and assume that a is the last record value prior to t. That means that $Z + Y(t-) < a$, $Z + Y(t) > a$ and by the lack of memory property of the exponential jumps $(Z + Y(t) - a) \backsim \exp(\mu)$. Hence, for every $x > a$, the failure rate function of that record value at x (that occurred at t) is μ and the event $\{M \leq x\}$ (which means that $\{M = x\}$) occurs if and only if the record value at t is the last record value in $[0, T_Z^-]$. The latter event occurs with probability $\Pr(T_x^- < T_x^+)$.

The argument introduced above is used as the main tool in proving theorem 5 below. But before we introduce the theorem we need the next two lemmas. These lemmas can also be retrieved from Section III.5.8 of Cohen [2], but we believe that the method of proof that is presented below is of independent interest.

Lemma 3 *Let* $\mathbf{V}_{M/G/1} = \{V_{M/G/1}(t) : t \geq 0\}$ *be the work process of the* $M/G/1$ *queue with arrival rate* μ *and service time distribution* $G(\cdot)$ *and assume that* $\rho := \lambda/\mu > 1$*. Given that the first service in the busy period is* a*, we define* $\theta_a(0, a+x)$ *as the probability that during a busy period* $\mathbf{V}_{M/G/1}$ *reaches level* 0 *before level* $a+x$ *(for* $x = 0$*, one should read here* $\theta_a(0, a+)$*). Then*

$$\theta_a(0, a + x) = \frac{F(x)}{F(a + x)}, \tag{25}$$

where $F(\cdot)$ *is the steady state distribution of* $\mathbf{V}_{M/G/1}$*. That is,* $F(\cdot)$ *is the distribution whose LST is given by*

$$F^*(\alpha) = \frac{(1 - \rho^{-1})\alpha}{\alpha - \mu[1 - G^*(\alpha)]}. \tag{26}$$

Proof First note that $\rho > 1$ implies that $\mathbf{V}_{M/G/1}$ possesses a stationary distribution. Recall that $-X(t) := \tilde{X}(t) = S_1 + S_2 + ... + S_{N(t)} - t$ where $N(t)$ is a Poisson process with rate μ. Also let $L_a = \inf\{t > 0 : a+\tilde{X}(t) = 0\}$ and $\tilde{M}_a = \max_{0\le t\le L_a}\{a + \tilde{X}(t)\}$. L_a can be interpreted as the busy period and $\tilde{M}_a$ as the cycle maximum of the VWT in the $M/G/1$ queue given that the first service of that busy period is a. Then

$$\begin{aligned} F(x) &= \Pr(\max_{0\le t<\infty} \tilde{X}(t) \le x) \\ &= \Pr(\max_{0\le t<\infty}(a + \tilde{X}(t)) \le a + x) \\ &= \Pr(\max_{0\le t<T_a}(a + \tilde{X}(t)) \le a + x)\Pr(\max_{T_a\le t<\infty}(a + \tilde{X}(t)) \le a + x) \\ &= \Pr(\tilde{M}_a \le a + x)F(a + x) \\ &= \theta_a(0, a + x)F(a + x). \end{aligned}$$

In particular, it follows by Lemma 3 that

$$\theta_a(0, a) = \frac{1 - \rho^{-1}}{F(a)}. \tag{27}$$

Lemma 4 below is based on the duality between the $M/G/1$ and the $G/M/1$ queues. Consider the VWT of the $G/M/1$ queue with inter-arrival distribution $G(\cdot)$ and service rate μ in which the first service of the busy period is x. Also, consider the VWT of the $M/G/1$ queue with arrival rate μ and service distribution $G(\cdot)$ in which the first service of the busy period is x.

Lemma 4

$$\Pr(T_x^- < T_x^+) = 1 - \frac{F * G(x)}{F(x)}$$

where the LST associated with $F(\cdot)$ is given in (26).

Proof Consider a sample path of the stopped process $\{x + Y(t) : 0 \le t \le T_x^-\}$ (see Fig. 1(*a*)). This stopped process represents the VWT of a $G/M/1$ queue during a busy period whose first service time is x. Now construct the risk stopped process $\{R(t) : 0 \le t \le T_x^-\}$ where $R(t) = -Y(t)$ (see Fig. 1(b)). That is, $R(t)$ starts at level 0 and is stopped immediately after it upcrosses level x. Now construct the process $\mathbf{U} = \{U(t) : t \ge 0\}$ from $\{R(t) : 0 \le t \le T_x^-\}$ as follows: First, replace every negative jump in Fig. 1(*b*) by a linearly decreasing piece of trajectory with slope -1 on an interval whose length is equal to the negative jump size. Second, replace the increasing pieces of $R(t)$ between negative jumps by positive jumps whose sizes are equal to the linear increments (the process is shown in Fig. 1(*c*)).

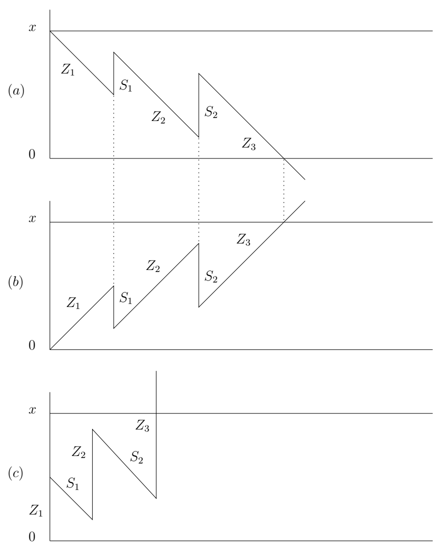

Fig. 1 The stopped process $\{x+Y(t) : 0 \le t \le T_x^-\}$ in (a), the risk stopped process $\{R(t) : 0 \le t \le T_x^-\}$ in (b) and the process $\mathbf{U} = \{U(t) : t \ge 0\}$ in (c).

Clearly, the event $\{T_x^- < T_x^+\}$ occurs if and only if level 0 is not downcrossed before level x is upcrossed by $\{x+Y(t) : 0 \le t \le T_x^-\}$. By construction of $R(t)$ the latter event occurs if and only if level x is upcrossed before level 0 is downcrossed by $\{R(t) : 0 \le t \le T_x^-\}$. Finally, by construction the latter event occurs if and only if level x is upcrossed before level 0 is downcrossed by $\mathbf{U}$. By this duality it can be seen that $\mathbf{U}$ represents the work process of the $M/G/1$ queue until the first upcrossing above level x (see also [7]). In order for the process $\mathbf{U}$ to upcross level x before downcrossing level 0 we condition on the size of the first jump. If the first jump is greater than x, level x is upcrossed at time 0. The latter event occurs with probability $1-G(x)$. If the first jump is $a < x$, then level x is upcrossed before level 0 is downcrossed with probability $1-\theta_a(0,x)$. Applying this argument we obtain

$$\Pr(T_x^- < T_x^+) = 1 - G(x) + \int_0^x [1-\theta_a(0,x)]dG(a). \tag{28}$$

Now replace x by $x-a$ in (25) and substitute in (28). The proof is complete after some elementary algebra.

We are now in a position to introduce the main result of this section.

Theorem 5

$$\Pr(M \le x | T < \infty) = \rho[1 - e^{-\mu\int_0^x (1-\frac{F*G(y)}{F(y)})dy}] = \rho - \frac{\rho-1}{F(x)}.$$

Proof The jumps of the VWT in the $G/M/1$ queue are $\exp(\mu)$. As mentioned before, μdx is the infinitesimal probability that an arbitrary record value of the VWT lands in $[x, x+dx)$. But x is the maximum of the VWT if and only if the latter record value is the last record value in the busy period, and the probability of that event is $\Pr(T_x^- < T_x^+)$. Multiplying, we conclude that the hazard rate function of M is

$$r(x) = \mu \Pr(T_x^- < T_x^+).$$

The first result of the theorem follows by Lemma 4, also observing that $\Pr(T < \infty) = 1/\rho$ (cf. [2], p. 217). The second result of the theorem is obtained as follows. Consider the steady-state work process in the $M/G/1$ queue with arrival rate μ and service time distribution $G(\cdot)$ (this steady-state law exists since $1/\rho = \mu/\lambda < 1$). It follows from the integro-differential equation of Takacs for the $M/G/1$ work process (cf. [2], p. 263) that

$$f(x) = \mu[F(x) - F * G(x)], \quad x > 0,$$

where F is the steady-state law of the work process and $f(x)$ is the density of $F(x)$, $x > 0$. We can now write:

$$\Pr(M \le x | T < \infty) = \rho[1 - \mathrm{e}^{-\int_0^x \frac{f(y)}{F(y)} \mathrm{d}y}] = \rho[1 - \mathrm{e}^{-\ln F(x) + D}].$$

The result follows by normalization.

Remark 10 It should be observed, using (27), that $\Pr(M \le x | T < \infty) = \rho[1 - \theta_x(0, x)]$, or $\Pr(M > x) = \theta_x(0, x)$. The latter result also follows from the construction in Figure 1.

References

1. S. Asmussen, *Applied Probability and Queues* (2nd edition, Springer-Verlag, New York 2003).
2. J.W. Cohen, *The Single Server Queue* (North-Holland Publ. Cy., Amsterdam 1982).
3. B.T. Doshi, A note on stochastic decomposition in a $GI/G/1$ queue with vacations or set-up times. J. Appl. Probab. **22** (1985), 419-428.
4. O. Kella and W. Whitt, Useful martingales for stochastic storage processes with Lévy input. J. Appl. Probab. **29** (1992), 396-403.
5. R. Núñez Queija, Note on the GI/GI/1 queue with LCFS-PR observed at arbitrary times. Prob. Eng. Inf. Sci. **15** (2001), 179-187.
6. D. Perry, W. Stadje and S. Zacks, Contributions to the theory of first-exit times of some compound processes in queueing theory. Queueing Systems **33** (1999), 369-379.
7. D. Perry and W. Stadje, Duality of dams via mountain processes. Operations Research Letters **31** (2003), 451-458.
8. D. Perry, W. Stadje and S. Zacks, Busy period analysis for $M/G/1$ and $G/M/1$-type queues with restricted accessibility. Operations Research Letters **27** (2000), 163-174.

9. N.U. Prabhu, *Queues and Inventories* (Wiley, New York 1965).
10. N.U. Prabhu, *Stochastic Storage Processes* (Springer-Verlag, New York 1988).
11. L. Takács, *Introduction to the Theory of Queues* (Oxford University Press, New York 1962).
12. E.C. Titchmarsh, *The Theory of Functions* (2nd edition; Oxford University Press, London 1968).